Available online at www.sciencedirect.com

SCIENCE DIRECT

Virology 348 (2006) 57–71

VIROLOGY

www.elsevier.com/locate/yviro

Complete nucleotide sequence and genome analysis of bacteriophage BFK20—A lytic phage of the industrial producer *Brevibacterium flavum*

Gabriela Bukovska [a,*,1], Lubos Klucar [a,1], Cestmir Vlcek [b], Jan Adamovic [a], Jan Turna [c], Jozef Timko [a]

[a] *Institute of Molecular Biology, Centre of Excellence for Molecular Medicine, Slovak Academy of Sciences, Dubravska cesta 21, 845 51 Bratislava, Slovakia*

[b] *Institute of Molecular Genetics, Academy of Sciences of the Czech Republic and Centre for Applied Genomics, Flemingovo nam. 2, Prague 6, CZ-16637, Czech Republic*

[c] *Comenius University, Faculty of Natural Sciences, Department of Molecular Biology, Mlynska dolina B-2, 842 15 Bratislava 4, Slovakia*

Received 22 July 2005; returned to author for revision 14 November 2005; accepted 11 December 2005
Available online 2 February 2006

Abstract

The entire double-stranded DNA genome of bacteriophage BFK20, a lytic phage of the *Brevibacterium flavum* CCM 251 – industrial producer of L-lysine – was sequenced and analyzed. It consists of 42,968 base pairs with an overall molar G + C content of 56.2%. Fifty-five potential open reading frames were identified and annotated using various bioinformatics tools. Clusters of functionally related putative genes were defined (structural, lytic, replication and regulatory). To verify the annotation of structural proteins, they were resolved by 2D gel electrophoresis and were submitted to N-terminal amino acid sequencing. Structural proteins identified included the portal and major and minor tail proteins. Based on the overall genome sequence comparison, similarities with other known bacteriophage genomes include primarily bacteriophages from *Mycobacterium* spp. and some regions of *Corynebacterium* spp. genomes – possible prophages. Our results support the theory that phage genomes are mosaics with respect to each other.

Keywords: Corynebacterium; Siphoviridae; Bacteriophage; BFK20 genome; Lytic phage; Complete genome sequence; Sequence analysis; 2D gel electrophoresis

Introduction

The genera *Corynebacterium* and *Brevibacterium* are Gram-positive bacteria of high G + C content, close to *Nocardia* and *Mycobacteria* and not very distant from *Streptomyces* (Park et al., 1987; Woese, 1987). The non-pathogenic members of genus *Corynebacterium* are widely used as industrial producers in many biotechnological processes (Batt et al., 1985; Krämer, 1996; Nampoothiri and Pandey, 1998). Applications range from feed to food and pharmaceutical products. *C. glutamicum* and *Brevibacterium flavum* play a major nearly exclusive role in the production of economically important amino acids (Hermann, 2003), including L-glutamate, threonine, L-lysine, phenylalanine, glutamine, arginine, tryptophan, isoleucine and histidine.

A serious and still common problem in many biotechnological laboratories and factories is infections of bacterial cultures by bacteriophages. The existence of lytic and temperate phages in corynebacteria has been known for many years (Kato et al., 1984, Patek et al., 1985). Most characterized corynephages are temperate and were isolated after UV irradiation or mitomycin induction (Moreau et al., 1995), but there are not many lytic corynephages defined in detail to date.

Corynephages usually have a polyhedral head (40 × 50–120 × 50 nm) attached to a long, flexible and non-contractile tail (80–250 nm) (Nampoothiri and Pandey, 1998). They contain double-stranded DNA with a G + C content of 54–57%.

* Corresponding author. Fax: +421 2 59307416.
E-mail addresses: Gabriela.Bukovska@savba.sk (G. Bukovska), Lubos.Klucar@savba.sk (L. Klucar), vlcek@img.cas.cz (C. Vlcek), Jan.Adamovic@savba.sk (J. Adamovic), turna@fns.uniba.sk (J. Turna), Jozef.Timko@savba.sk (J. Timko).
[1] Both authors contributed equally.

0042-6822/$ - see front matter © 2005 Elsevier Inc. All rights reserved.
doi:10.1016/j.virol.2005.12.010

The host range, virion morphology, DNA and protein profiles of several corynephages have been reported (Trautwetter and Blanco, 1988; Trautwetter et al., 1987a). Phages characterized in detail include phages Φ304L, Φ304S, Φ15 and Φ16, which were induced from different *C. glutamicum* ATCC derivatives (Moreau et al., 1995, 1999), ΦGA1 – *B. flavum* phage (Sonnen et al., 1990a), β-converting and γ-nonconverting corynebacteriophage of the pathogenic species *C. diphtheriae* (Buck and Groman, 1981) as well as ΦAAU2 infecting *"Arthrobacter aureus"-C70* (Le Marrec et al., 1996). Bacteriophage CL31 infects *C. lilium* (Trautwetter et al., 1987b), and Cog phage is a virulent phage of *C. glutamicum* (Sonnen et al., 1990b). According to Ackermann (2003), over 5300 phages have been examined by electron microscopy. In the EMBL database, we could find DNA segments from around 1000 different bacteriophages, giving the approximate number of known bacteriophages studied at a molecular level. There are more than 250 completely known genomes of bacteriophages, but no complete nucleotide sequence of any corynephage has yet been reported.

We have previously isolated and characterized corynephage BFK20, which causes lysis of *B. flavum* CCM 251, the industrial producer of L-lysine (Koptides et al., 1992, 1994). Previous study of the host range of BFK20 (Koptides et al., 1992) revealed only one host strain lysed after infection. No lysis of other *B. flavum* ATCC strains, *B. lactofermentum* BLOB and *C. glutamicum* ATCC 13032 was observed. Defense mechanisms of corynebacteria strains against bacteriophage BFK20 were recently investigated in detail (Halgasova et al., 2005).

B. flavum is considered to be the same species as *C. glutamicum* (Liebl et al., 1991). Thus, known genomic sequence of *C. glutamicum* ATCC 13032 (Ikeda and Nakagawa, 2003; Kalinowski et al., 2003) was a great benefit for the BFK20 bioinformatic analysis presented in this study. Bacteriophage BFK20 morphologically belongs to a taxonomical group of unclassified Siphoviridae (Koptides et al., 1992). The phage particle is composed of 50 nm polyhedral head and 200 × 10 nm non-contractible tail. The genome of BFK20 contains a linear double-stranded DNA molecule with 3′ cohesive ends and a GC content of 56.2% (EMBL, accession no. AJ278322). The BFK20 genome consists of 42,968 bp, and it is the first corynephage to be completely sequenced. Using bioinformatics analysis, we have identified 55 putative open reading frames (ORFs) coding for proteins varying in molecular weight (Mw) from 5 to 170 kDa.

Based on the overall genome sequence comparison, similarities with other known bacteriophage genomes were found. Those primarily include bacteriophages from *Mycobacterium* spp. host group and some regions of *Corynebacterium* spp. genomes – possible prophages. Protein-based homology search revealed similarities between BFK20 proteins and other phage and bacterial proteins, which helped predict their functions. Analyses of both DNA and protein support the theory that phage genomes are mosaics with respect to each other. This makes phylogenetic relations of phages more non-linear in comparison to other organisms.

To date, most of the completed genome sequences exist for the phages that infect Gram-negative eubacteria or A + T-rich Gram-positive bacteria such as *Lactococci*, *Streptococci* or *Bacilli*. In this study, we present the complete annotated sequence of the BFK20 genome – a lytic phage of the Gram-positive industrial producer *B. flavum* – and analyze the predicted virion proteins.

Results and discussion

Determination of DNA sequence

The nucleotide sequence of the BFK20 genome was determined by two approaches. Initially, genomic clones containing defined fragments were prepared according to the restriction map (Koptides et al., 1992). The complete nucleotide sequences of these clones were determined step by step using automatic DNA sequencers. Final localization and orientation of sequences on the phage genome were completed by primer walking. Secondly, the whole BFK20 genome was cloned by shotgun, and individual clones were sequenced. The whole nucleotide sequence of BFK20 DNA was assembled (EMBL Acc. No.: AJ278322) using Gap4 program from Staden Package (Staden, 1996). The genome size was estimated as 42,968 bp. Each nucleotide was determined at least twice and up to 10 times in both directions. The overall G + C content of the BFK20 genome is 56.2%. G + C content of *C. glutamicum*, which is considered to be the same species as BFK20 host *B. flavum* (Liebl et al., 1991), is 53.8%.

The presence of *cos* sites is characteristic for phages with a non-headful packaging mechanism. We suggested cohesive ends for the BFK20 genome according to previous results (Koptides et al., 1992). The exact *cos* site sequence was determined by sequencing on genomic DNA in a sequence run-off experiment using the oligonucleotides COS1 and COS2. Primers were situated 277 bp distal to either end of the region containing the expected *cos* site. By comparing these sequences with the previously assembled whole genome (ligated DNA fragments from the shotgun strategy), the *cos* site was assigned to a 13 bp sequence 3′-ACTTCCCCCGCTT and TGAAGGGGGCGAA-3′. Thus, the BFK20 *cos* site possesses a single-stranded 3′ overhang, like many other phages infecting Gram-positive organisms (Brüssow and Desiere, 2001, Chandry et al., 1994; Ganyu et al., 2005; Kaneko et al., 1998; Lillehaug et al., 1993; Mahanivong et al., 2001; van Sinderen et al., 1996).

ORFs prediction and genomic organization of bacteriophage BFK20

Identification of coding regions, located on the BFK20 genome, was based on the application of various methods. These include base preferences, codon preferences, author test, base bias and the presence of START and STOP codons (Staden, 1996). A DNA region was considered as coding if START and STOP codons were present, at least one of the base composition methods was positive and the length of ORF was at least 100 bp. All ORFs begin with ATG/GTG, except for ORFs 14 and 36 which use TTG. As summarized in Table 1, we identified 55 ORFs, predicted as coding regions. These 55

Table 1
Description of predicted ORFs and proteins from BFK20

Orf	Start	Stop	(G + C)3	Length [AA]	Mw [kDa]	p*I*	Predicted function	Related proteins and protein families	Tool	E value
Late										
1	126	554	48.3	142	15.0	9.7	Small terminase	N/A		
2	538	2166	54.2	542	61.1	4.5	Large terminase	BAD84098 - *C. glutamicum*	Blastp	3.0E-19
								gp38 - *Streptococcus mitis* phage SM1	Blastp	3.0E-08
								phage terminase - *Streptococcus pyogenes* MGAS10394	Blastp	4.0E-08
								gp2 - Mycobacteriophage Che9c	Blastp	4.0E-08
								Putative terminase gp4 - Mycobacteriophage TM4	Blastp	2.0E-05
								Putative terminase - Bacteriophage r1t	Blastp	3.0E-05
								gp11 - Mycobacteriophage Omega	Blastp	1.0E-04
								Phage terminase	HmmPfam	5.0E-05
								Lambda-repressor HTH signature	Pscan	N/A
3	2178	3449	60.2	423	47.7	4.8	Portal protein	DIP0203 – *C. diphtheriae* NCTC 13129	Blastp	1.0E-102
								nfa410 - *Nocardia farcinica* IFM 10152	Blastp	4.0E-56
								Phage portal protein gp10 - Mycobacteriophage CJW1	Blastp	1.0E-13
								Phage portal protein gp13 - Mycobacteriophage Omega	Blastp	3.0E-07
								Head portal protein - Enterobacteria phage HK022	Blastp	7.0E-06
								Portal protein gp3 - Bacteriophage HK97	Blastp	2.0E-05
								Phage portal protein	HmmPfam	3.1E-11
4[a]	3494	3775	56.4	93	9.9	6.8	N/A	Predicted P-loop containing kinase - *Thermobifida fusca*	Blastp	6.0E-06
								P-loop ATPase protein family	HmmPfam	1.1E-06
5	3772	4824	62.6	347	37.5	4.4	Prohead protease	Putative phage prohead protease - *C. diphtheriae* NCTC 13129	Blastp	3.0E-98
								Phage head maturation protease - *Silicibacter* sp. TM1040	Blastp	2.0E-22
								CE0839 - *C. efficiens* YS-314	Blastp	3.0E-10
								Prohead protease gp4 - Bacteriophage HK97	Blastp	4.0E-06
								Head maturation protease - Enterobacteria phage HK022	Blastp	7.0E-06
								Prohead protease gp14 - Mycobacteriophage Omega	Blastp	2.0E-04
								Caudovirus prohead protease	HmmPfam	2.4E-82
								Conserved hypothetical ATP binding protein	HmmPfam	9.2E-01
								ATP_GTP_A ATP/GTP-binding site motif A (P-loop)	PrositeSearch	N/A
6	4859	6100	47.3	413	45.1	4.6	Major capsid protein	Putative phage capsid protein - *C. diphtheriae* NCTC 13129	Blastp	9.0E-80
								Phage major capsid protein - Pseudomonas phage D3	Blastp	9.0E-16
								gp12 - Mycobacteriophage CJW1	Blastp	6.0E-12
								Major head subunit precursor - Bacteriophage HK97	Blastp	8.0E-09
								Major capsid subunit precursor - Enterobacteria phage HK022	Blastp	8.0E-09
								Phage capsid family	HmmPfam	1.2
								Bacteriophage lambda head decoration	HmmPfam	3.0
7	6100	6303	52.4	67	7.2	11.1	N/A	N/A		
8	6306	6848	53.8	180	19.9	6.6	Structural protein (exp.[b])	DIP0207 - *C. diphtheriae* NCTC	Blastp	2.0E-20
								gp25 - Mycobacteriophage Omega	Blastp	3.0E-07
								gp8 - Mycobacteriophage Che9c	Blastp	1.6E-02
								gp43 - Mycobacteriophage Corndog	Blastp	2.8E-02
9	6860	7216	62.9	118	13.1	4.2	N/A	DIP0208 - *C. diphtheriae* NCTC	Blastp	2.0E-18
10	7209	7493	47.5	94	10.2	12.0	N/A	DIP0208A - *C. diphtheriae* NCTC	Blastp	5.0E-06
11	7541	7870	54.5	109	11.8	6.8	N/A	DIP0209 - *C. diphtheriae* NCTC	Blastp	1.0E-07

(continued on next page)

Table 1 (*continued*)

Orf	Start	Stop	(G + C)3	Length [AA]	Mw [kDa]	p*I*	Predicted function	Related proteins and protein families	Tool	E value
Late										
12	7886	8830	58.2	314	33.4	4.5	Major tail protein	DIP0210 - *C. diphtheriae* NCTC	Blastp	3.0E-57
						14.5		Putative phage major tail protein - *Streptomyces avermitilis* MA–4680	Blastp	3.0E-08
								Majortail protein gp14 - Mycobacteriophage Che9d	Blastp	3.7E-02
								gp15 - Mycobacteriophage Rosebush	Blastp	4.1E-01
								gp20 - Mycobacteriophage Bethlehem	Blastp	9.1E-01
								gp20 - Mycobacteriophage U2	Blastp	9.1E-01
13	8927	9343	50.0	138	15.7	4.9	N/A	DIP0211 - *C. diphtheriae* NCTC	Blastp	1.0E-10
14	9427	9702	50.0	91	10.4	7.5	N/A	DIP0211A - *C. diphtheriae* NCTC	Blastp	4.7
15	9713	14,527	52.2	1604	168.9	4.3	Minor tail protein (tape measure protein)	BAD84085 - *C. glutamicum*	Blastp	0
								CE0854 - *C. efficiens* YS-314	Blastp	1.0E-135
								ORF9 - Bacteriophage VWB	Blastp	3.0E-76
								DIP0212 - *C. diphtheriae* NCTC	Blastp	8.0E-38
								Phage-related minor tail protein - *Enterococcus faecium*	Blastp	4.0E-32
								gp15 - Mycobacteriophage Che9c	Blastp	1.0E-20
								PblA - *Streptococcus mitis* phage SM1	Blastp	1.0E-20
								Putative tape-measure protein - Mycobacteriophage TM4	Blastp	2.0E-10
								gp28 - Mycobacteriophage PG1	Blastp	4.0E-05
								gp29 - Mycobacteriophage Rosebush	Blastp	6.0E-03
								gp133 - Mycobacteriophage Bxz1	Blastp	2.1E-01
								WD_REPEATS_1 Trp-Asp (WD) repeats signature	PrositeSearch	N/A
								Lambda-repressor HTH signature	Pscan	N/A
								Transglycosylase SLT domain	HmmPfam	4.3E-16
								Prophage tail length tape measure protein	HmmPfam	2.0E-01
16	14,538	15,356	62.4	272	29.1	4.7	Structural protein (exp.)	BAD84084 - *C. glutamicum*	Blastp	6.0E-09
								CE0855 - *C. efficiens* YS-314	Blastp	8.6E-02
								Tail assembly protein gp18 – Mycobacteriophage TM4	Psi-Blastp	2.4E-01
17	15,361	15,960	54.5	199	22.0	5.7	N/A	N/A		
18	15,932	16,987	56.4	351	38.3	6.4	Tail assoc. protein	CE0856 - *C. efficiens* YS-314	Blastp	2.0E-40
								BAD84083 - *C. glutamicum*	Blastp	3.0E-31
								Putative phage tail - *Nocardia farcinica* IFM 10152	Blastp	1.0E-15
								gp2 - Mycobacteriophage CJW1	Blastp	5.0E-14
								gp36 - Mycobacteriophage Omega	Blastp	3.0E-12
19	16,998	17,381	55.0	127	13.6	4.1	N/A	BAD84082 - *C. glutamicum*	Blastp	3.0E-03
								CE0857 - *C. efficiens* YS-314	Blastp	1.1E-01
20	17,371	18,600	52.3	409	44.3	4.9	N/A	BAD84081 – *C. glutamicum*	Blastp	3.0E-79
21	18,613	19,470	53.0	285	28.4	7.0	Tail fiber protein	L-shaped tail fiber protein - phage T5	Blastp	4.0E-04
								CE0859 - *C. efficiens* YS-314	Blastp	5.0E-14
								ALA_RICH Alanine-rich region	PrositeSearch	N/A
22	19,482	20,558	56.4	358	37.3	6.7	N/A	gp232 - Mycobacteriophage Bxz1	Blastp	9.0E-19
23	20,561	20,905	40.3	114	12.1	8.5	N/A	N/A		
24	20,961	21,536	60.2	191	21.7	10.6	Lysin	BAD84076 - *C. glutamicum*	Blastp	2.0E-77
								Autolysin - *Staphylococcus aureus*	Blastp	9.0E-05
								CE0861 - *C. efficiens* YS-314	Blastp	4.0E-14
								N-acetylmuramoyl-L-alanine amidase	HmmPfam	1.2E-03
25	21,524	21,772	62.5	82	9.3	9.8	N/A	BAD84076 - *C. glutamicum*	Blastp	2.0E-18
								DIP0218 - *C. diphtheriae* NCTC 13129	Blastp	6.0E-05

26	21,772	22,323	60.5	183	19.5	4.3	Holin (exp.)	BAD84075 - *C. glutamicum* 2-phosphosulpholactate phosphatase	Blastp HmmPfam	5.0E-05 1.7
27	22,371	22,790	60.9	139	15.1	5.6	N/A	AD84074 - *C. glutamicum* Class II bacteriocin	Blastp HmmPfam	4.0E-09 1.9
28	22,798	23,214	62.1	138	15.9	7.7	N/A	GGDEF domain containing protein - *Rubrivivax gelatinosus* PM1	Blastp	2.3E-02
Early										
29	23,648	23,265	57.8	127	15.0	10.4	N/A	gp56 - Mycobacteriophage Bxz1 Putative endonuclease VII - *Streptomyces avermitilis* MA-4680 gp199 – Mycobacteriophage Omega Lambda-repressor HTH signature	Blastp Psi-blast Psi-blast Pscan	3.0E-02 2.0E-22 3.0E-15 N/A
30	23,937	23,689	65.0	82	9.1	4.0	N/A	N/A		
31	24,211	23,990	62.5	73	8.3	4.4	N/A	N/A		
32	24,537	24,304	57.4	77	9.0	8.9	Transcriptional regulator	MerR family regulatory protein	HmmPfam	3.6E-02
33	24,945	24,646	56.1	99	11.3	4.6	N/A	SH2 domain	HmmPfam	2.5E-01
34	25,497	25,294	57.1	67	7.6	7.8	N/A	N/A		
35	25,898	25,566	72.3	110	12.5	5.2	N/A	N/A		
36	26,690	26,412	56.6	92	10.3	3.8	N/A	N/A		
37	27,577	26,879	47.8	232	26.1	4.3	Transcriptional regulator	gp1.7 - *Yersinia pestis* phage phiA1122 Lambda-repressor HTH signature Adenylate cyclase associated	Blastp Pscan HmmPfam	1.0E-04 N/A 1.1E-01
38	28,089	27,808	38.2	93	9.9	7.1	N/A	Exodeoxyribonuclease V, beta subunit - *Pseudomonas syringae* pv. Tomato str. DC3000	Blastp	2.1
39	29,287	28,406	58.4	293	32.3	4.2	N/A	GLN_RICH Glutamine-rich region*O*	PrositeSearch	N/A
40	30,566	29,511	58.1	351	39.2	5.2	N/A	gp49 - Mycobacteriophage Rosebush gp52 - Mycobacteriophage PG1 gp68 - Mycobacteriophage CJW1	Blastp Blastp Blastp	3.0E-36 3.0E-32 1.0E-03
41	32,248	30,635	49.2	537	57.9	5.8	Helicase	Helicase (gp50) - Mycobacteriophage Rosebush gp53 - Mycobacteriophage PG1 DNA or RNA helicases of superfamily II - *C. glutamicum* ATCC 13032 Putative helicase – *C. diphtheriae* NCTC 13129 Putative DNA helicase - *C. efficiens* YS-314 ATP_GTP_A ATP/GTP-binding site motif A (P-loop) DEAD/DEAH box helicase Helicase conserved C-terminal domain	Blastp Blastp Blastp Blastp Blastp PrositeSearch HmmPfam HmmPfam	6.0E-98 2.0E-96 2.0E-27 1.0E-26 2.0E-09 N/A 1.1E-04 7.9E-18
42	32,595	32,248	58.8	115	12.3	4.1	N/A	Hypothetical protein - *C. glutamicum* ATCC 13032 gp173 - Mycobacteriophage Omega gp113 - Mycobacteriophage CJW1*T*	Blastp Blastp Blastp	4.0E-14 2.0E-11 1.0E-07
43	36,016	33,068	57.1	982	107.7	4.9	DNA replication protein	gp59 - Mycobacteriophage PG1 gp54 - Mycobacteriophage Rosebush RepA - *Aeromonas salmonicida* ATP_GTP_A ATP/GTP-binding site motif A (P-loop) ABC transporter	Blastp Blastp Blastp PrositeSearch HmmPfam	2.0E-68 6.0E-62 1.0E-08 N/A 2.6E-02

(continued on next page)

Table 1 (*continued*)

Orf	Start	Stop	(G + C)3	Length [AA]	Mw [kDa]	p*I*	Predicted function	Related proteins and protein families	Tool	E value
44	37,988	36,084	60.3	634	70.2	6.4	DNA polumerase A	gp56 - Mycobacteriophage Rosebush	Blastp	1.0E-119
								gp61 - Mycobacteriophage PG1	Blastp	1.0E-105
								DNA polymerase I - *Caulobacter crescentus* CB15	Blastp	9.0E-21
								DNA polymerase I - *C. glutamicum* ATCC 13032T	Blastp	2.0E-13
								Putative DNA polymerase I - *C. efficiens* YS-314	Blastp	9.0E-13
								DNA polymerase I - *C. diphtheriae* NCTC 13129T	Blastp	1.0E-12
								DNA_POLYMERASE_A DNA polymerase family A signature	PrositeSearch	N/A
								DNA-pol I	FPScan	6.9E-01
								DNA polymerase family A	HmmPfam	1.2E-02
45	38,639	38,160	70.8	159	18.2	5.3	N/A	N/A		
46	38,973	38,641	57.4	110	12.5	7.1	N/A	Putative methyltransferase	HmmPfam	3.2
47	39,392	38,994	51.3	132	14.7	5.5	Transcriptional regulator	Transcriptional regulator	HmmPfam	1.7
48	39,777	39,385	59.1	130	14.2	10.3	Transcriptional regulator	Bacterial regulatory proteins, gntR family	HmmPfam	1.8E-01
49	40,273	40,034	45.2	79	8.6	7.9	N/A	N/A		
Late										
50	40,857	41,033	61.5	58	6.6	9.6	N/A	N/A		
51	41,030	41,161	52.2	43	4.7	3.8	N/A	N/A		
52	41,168	41,554	59.6	128	14.3	6.7	Transcriptional regulator	Lactococcus bacteriophage putative transcription regulator	HmmPfam	2.6E-01
53	41,565	41,831	41.7	88	9.4	8.1	Glutaredoxin	Glutaredoxin protein - *Agrobacterium tumefaciens* str. C58	Blastp	1.0E-09
								Putative glutaredoxin - *Nocardia farcinica* IFM 10152	Blastp	6.0E-09
								Glutaredoxin and related proteins - *C. glutamicum* ATCC 13032	Blastp	1.0E-08
								Glutaredoxin gp37 - Mycobacteriophage CJW1	Blastp	2.0E-07
								Glutaredoxin gp56 - Mycobacteriophage L5	Blastp	6.0E-06
								Glutaredoxin gp56 - Mycobacteriophage D29	Blastp	2.0E-05
								Glutaredoxin	FPScan	4.9E-02
								Glutaredoxin	HmmPfam	1.0E-02
54	41,912	42,481	51.0	189	20.6	7.3	Transcriptional regulator	Lambda-repressor HTH signature	Pscan	N/A
								Acetohydroxy acid isomeroreductase, catalytic domain	HmmPfam	1.8E-01
55	42,586	42,876	57.8	96	10.5	5.0	HNH endonuclease	DIP0200 - *C. diphtheriae* NCTC 13129T	Blastp	1.0E-18
								HNH endonuclease family protein - *C. diphtheriaeT*	Blastp	1.1E-01
								HNH endonuclease	HmmPfam	7.4E-03

[a] Gray highlight indicates ORFs shorter than 100 codons.
[b] Exp. — experimentally detected.

ORFs represent 89% of the total genome. The average ORF length is 696 bp and intergenic regions 107 bp. Genes vary in length from 132 bp (ORF51) to 4815 bp (ORF15), 17 out of 55 predicted genes are shorter than 300 bp (Table 1). According to their direction of transcription, ORFs were divided into two groups. ORF1–ORF28, ORF50–ORF55 are transcribed from one strand and ORF29–ORF49 from the complementary strand (Fig. 1). The overall gene organization consists of at least two different classes of transcription: 1. early (DNA replication) and 2. late (structural proteins synthesis, maturation, lysis) (Fig. 1). Transcription of both groups points inwards, to position 23,250 bp. In this region, a bi-directional transcriptional termination loop is predicted. The terminator consists of a 30 bp hairpin loop (23,229–23,259) with a 10 bp long stem of 80% G + C content.

BFK20 contains 11% of non-coding sequence; in many cases, there is little or no space between genes. Three ORFs overlap by one nucleotide position where two adjacent genes contain *TAATG* (ORF6–ORF7, ORF25–ORF26, ORF42–ORF41). Two other ORFs contain the sequence *ATGA*, in which the START codon of the downstream gene overlaps the TGA STOP codon of the preceding gene (ORF4–ORF5, ORF50–ORF51). Six gene pairs overlap for between 7 and 28 nucleotides (ORF1–ORF2, ORF9–ORF10, ORF17–ORF18, ORF19–ORF20, ORF24–ORF25, ORF48–ORF47). This indicates that translation of some ORFs is coupled or an opportunity for a programmed translational frameshift is provided. However, no translational frameshift was detected between genes ORF13 and ORF14, which precede the minor tail protein gene. Translational frameshifting is common in this region among many tailed bacteriophages (Xu et al., 2004). We predicted a Shine-Dalgarno sequence (SD) for 44 out of the 55 BFK20 ORFs. The most frequent SD motifs are AGGAG, GGGGG, AGGGG and AGGCG. In *C. glutamicum*, the anti-Shine-Dalgarno sequence comprises a tract of nine pyrimidines with the sequence 5′-CCUCCUUUC at the 3′ end of the 16S rRNA (Martin et al., 2003).

The G + C content of coding regions is 56.5%, and (G + C)3 content (G + C content at the silent third position of sense codons) is 55.7%. These values are almost identical for both groups of ORFs (early genes — 57.0%, late genes — 56.2%). However, we observed great differences among ORFs. Since the (G + C)3 value has greater significance when comparing G + C content in coding DNA sequences, we discuss only these values. If we exclude very short (statistically insignificant) ORFs less than 300 bp long, we find that ORF23 exhibits extremely low (G + C)3 content — 40.3%. On the other hand, ORF35 and ORF45 possess extremely high (G + C)3 content — 72.3% and 70.8%, respectively. None of these gene

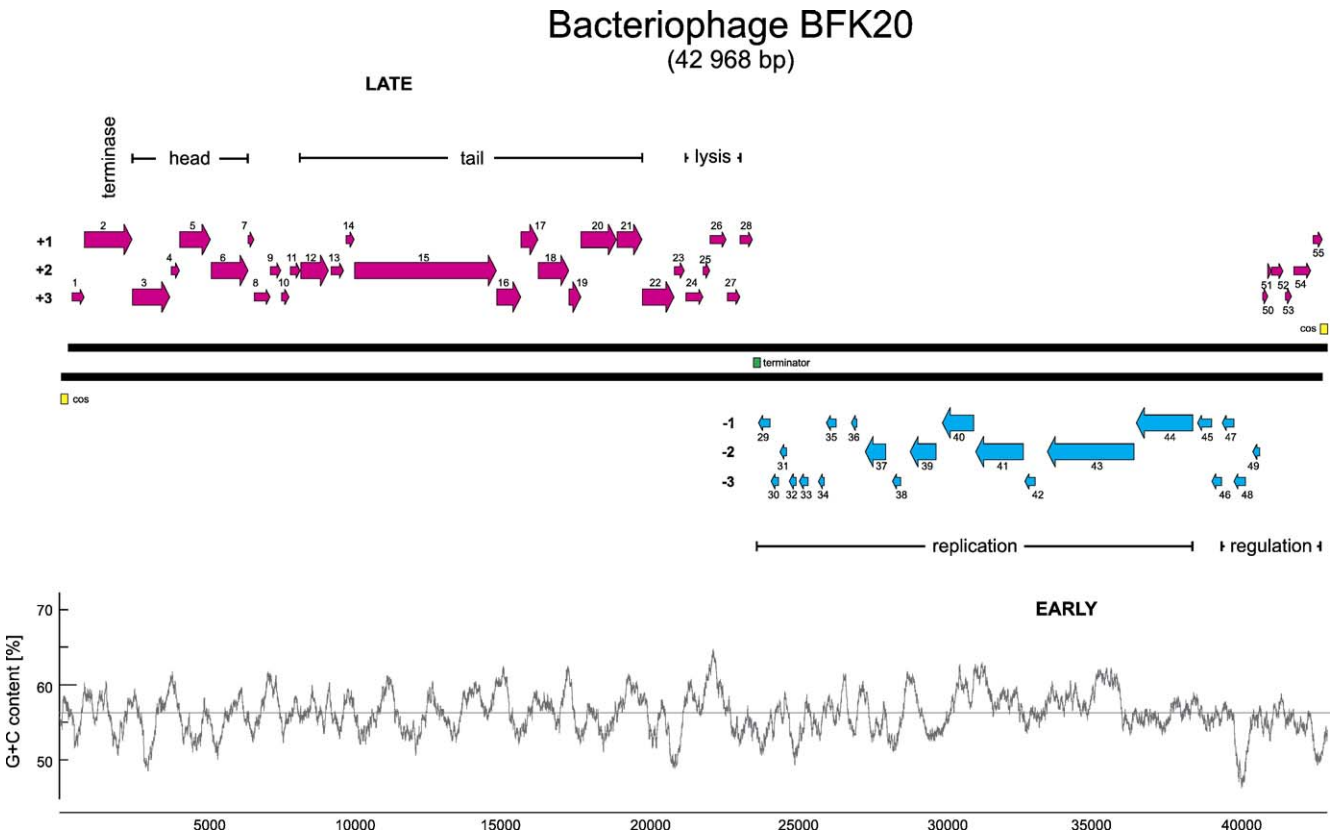


Fig. 1. Genomic organization of bacteriophage BFK20. The 42,968 bp BFK20 genome is represented by thick double line. Predicted genes are represented as colored arrows. The numbers +1, +2 and +3 represent corresponding reading frame. Early genes region consist of replication and regulation genes. The late genes region consists of terminase, head and tail components and lysis genes. Predicted bi-directional terminator at position ~23,250 bp is depicted as a green box. Yellow boxes represent 13 bp long 3′ protruding *cos* ends (3′-ACTTCCCCCGCTT and TGAAGGGGGCGAA-3′). G + C content profile is shown in lower half of the picture.

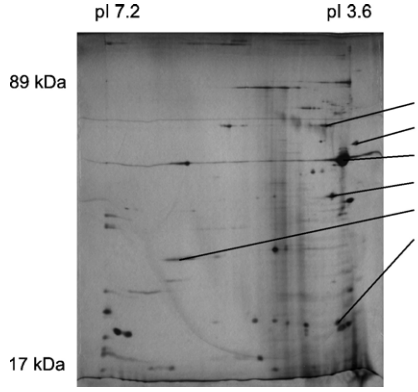
products was identified as proteins with specific functions and may represent pseudogenes or mobile elements without a coding function.

Codon usage analysis compared codon usage between the bacteriophage BFK20 and the bacterium *C. glutamicum* ATCC 13032. The distribution of codons used in BFK20 is very similar to the one used by *C. glutamicum*. All codon frequencies in the phage and bacterium were within 20% difference with only two exceptions: 1. both Glu codons (GAA, GAG) were used in reverse preference, and 2. the CCA Pro codon was used more frequently in BFK20, compared to *C. glutamicum* (Supplement 1). Genes that deviate strongly in codon usage from an average gene, but are relatively similar in a codon usage to the ribosomal protein genes, are considered as highly expressed (Karlin et al., 1998a, 1998b). Genes for 53 ribosomal proteins from *C. glutamicum* ATCC 13032 were collected from the EMBL database, and their codon usage profile was used as a standard codon usage pattern for highly expressed genes. We computed "B" values (Karlin et al., 1998b) and CAI (Codon Adaptation Index) (Sharp and Li, 1987) for 39 ORFs from BFK20 and for 2746 known genes from *C. glutamicum* ATCC 13032, considering only ORFs longer than 300 bp. The B value was used to compare the codon bias between BFK20 ORFs and average *C. glutamicum* ATCC 13032 genes. The CAI value was computed as a codon usage bias using the ribosomal proteins genes from *C. glutamicum* as a reference set. Based on the plot of B and CAI values, we estimate that the great majority of ORFs from BFK20 fall within the average codon usage of *C. glutamicum* (data not shown). Only ORF6, ORF33 and ORF39 were considered as highly expressed, according to their similarity in codon usage with ribosomal proteins and their dissimilarity with the average codon usage of *Corynebacterium glutamicum* ATCC 13032.

Proteomic analysis of BFK20 virion

Proteomic analysis of the BFK20 structural proteins was undertaken to aid the genome annotation. BFK20 phage was purified by CsCl density gradient centrifugation, and the virion proteins were analyzed by 2D gel electrophoresis (Fig. 2). Twenty three proteins spots were identified with different Mw mobilities. However, some of these exhibited different p*I* mobilities and were not resolved into a single discrete spot, suggesting either multiple protein species and/or post-translational modifications. Only thirteen proteins spots contained a sufficient amount of material for N-terminal amino acid sequencing. The first five amino acids for each protein were determined. Our analysis revealed six proteins in BFK20 genome that contained the corresponding amino acid sequences: gp3 (portal protein), gp5 (prohead protease), gp8, gp12 (major tail protein), gp15b (minor tail protein) and gp16. Except for gp5 and gp8, all proteins lack the initiating methionine, which is presumably due to the host methionine aminopeptidase activity (Lowther and Matthews, 2000). The estimated and predicted molecular weights of the analyzed proteins were different. All proteins, except for gp5, migrated with higher mobilities than expected. This could be due to retention of some secondary structure that influences SDS binding during electrophoresis. Similar observations have been noticed for other phage virion proteins (Labrie and Moineau, 2002; Roberts et al., 2004; Romero et al., 2004, Smith et al., 1999).

The significance of gp5 among structural proteins is unclear since it was identified as the prohead protease. The discrepancy between the experimentally estimated Mw of 16 kDa and the predicted 38 kDa for gp5 is also excessive. Based on our bioinformatic analysis (see below), we assumed a process where gp5 is matured into two 15 kDa and 23 kDa proteins. We detected sequence similarity between gp5 and known proteases starting only from amino acid position 140 of gp5. Therefore, the C-terminal part of gp5 could be the genuine protease and its predicted Mw of 23 kDa corresponds to those of other known proteases. The N-terminal part of gp5 protein (up to amino acid position 140) represents an unknown structural protein with a predicted Mw of 15 kDa. This corresponds to the protein with the N-terminal sequence MIHMI and an estimated Mw 16 kDa (Fig. 2). The HDGGL peptide (Fig. 2) is found at position 1299 AA of gp15. This could be a result of post-translational modification, in which protein gp15 is cleaved into a mature gp15a of Mw 138 kDa and gp15b of Mw 31 kDa. However, the experimentally estimated Mw for



protein product	N-terminal seq.	Mw [kDa] est.	Mw [kDa] pred.	pI est.	pI pred.	predicted function
gp3	GFLQK	59	48	4.9	4.8	portal protein
gp15b	HDGGL	48	31	4.1	4.0	minor tail protein
gp12	SNDYI	45	33	4.4	4.5	major tail protein
gp16	PGYRP	36	29	4.7	4.7	*unknown*
gp8	MQPPH	30	20	7.6	6.6	*unknown*
gp5	MIHMI	16	38	4.5	4.4	prohead protease

Fig. 2. 2D gel analysis of structural proteins. 2D electrophoretogram stained with Coomassie Blue R-250 is depicted on the left side. Maximal and minimal p*I* and Mw of standards are shown. Indicated proteins are those, which were determined using N-terminal sequencing. Mw and p*I* for each protein product were estimated from the 2D gel analysis (est.) and predicted from amino acid sequences (pred.). Obtained N-terminal sequence and predicted function for gene products are also shown. Gp15b represents hypothetical C-terminal part of gp15 (see Proteomic analysis of BFK20 virion).

gp15b of 48 kDa does not correspond to that expected for gp15b. Gp8 and gp16 are proteins with unknown function, but their position on the BFK20 genome and experimental confirmation of association to the group of phage virion proteins clearly indicated that both belong to the group of structural proteins.

Proteomic analysis of BFK20 virion particles did not detect the predominant protein in the phage particle — the major capsid protein, which was identified as gp6. Gp6 was identified as MCP with high similarity and homology to the group of HK97-like MCPs. A known feature of HK97 capsid assembly is the formation of crosslinks that link subunits in the capsid like a kind of chain mail. In bacteriophages containing chain mail MCPs, the major capsid protein is detected by SDS polyacrylamide gel electrophoresis only as a polypeptide with very high molecular weight that does not enter the gel (Gilakjan and Kropinski, 1999; Smith et al., 1999; Yuzenkova et al., 2003). The presence of chain mail is a very probable reason why BFK20 major capsid protein was not detected (Fig. 2). The SDS electrophoretic pattern of BFK20 virion proteins displayed the appearance of high molecular weight material that failed to enter the gel matrix (data not shown). This material probably contains a major capsid protein oligomer.

Characterization and function of predicted proteins

Many phages of the Siphoviridae family share a similar genome organization: "DNA packaging–head–tail-tail fiber-lysis/lysogeny-DNA replication-transcriptional regulation" modules (Brüssow and Desiere, 2001); BFK20 has an analogous structure. The 55 ORFs and their gene products were compared with the non-redundant databases using Blast (Altschul et al., 1997), and their associations to the protein families and other similarities were established using secondary databases search tools PrositeSearch (EMBOSS package software, Rice et al., 2000), Pscan (EMBOSS package software, Rice et al., 2000), FPScan (Scordis et al., 1999) and HmmPfam (Durbin et al., 1998). Descriptions of predicted proteins are summarized in Table 1.

Head morphogenesis and structural proteins gp1–gp22

In many tailed phages and prophages, the organization of head morphogenesis and structural proteins is: "terminase–portal–protease–scaffold–major head shell (coat) protein–head/tail-joining proteins–tail shaft protein–tape measure protein–tail tip/baseplate proteins–tail fiber" (listed in their order of transcription) (Casjens, 2003). In the BFK20 genome, the putative head morphogenesis and structural genes are located immediately upstream of the tail genes. The structural proteins of BFK20 were identified by bioinformatics and experimentally confirmed by the determination of N-terminal amino acid sequences of individual virion proteins.

Gp1 and gp2 are candidates for, respectively, the small and large terminase subunits. The terminase enzymes that have been characterized to date are mostly heteromultimers with a small subunit that specifically binds the viral DNA and a large subunit that has endonuclease activity for DNA cleavage and an ATPase activity that powers DNA packaging (Catalano et al., 1995, Duffy and Feiss, 2002; Maluf et al., 2005; Rubinchik et al., 1994). Gp1 has a similar size and genomic position (preceding the large terminase subunit) with known small terminase subunits (Brüssow and Desiere, 2001; Canchaya et al., 2003). Amino acid sequence alignment of gp1 with 99 known small terminase subunits was examined (data not shown); no significant sequence similarity was detected. This could be due to the fact that there is no conserved motif present among all known small terminase subunits. Evidence for terminase activity of gp2 is supported by several Blast matches to putative terminases from several other bacteriophages (Table 1). Terminase functionality was predicted also by an HmmPfam search. Sequence alignment of gp2 with other phage terminases revealed the presence of a highly conserved GKT/S Walker A (P-loop) motif VGRQN**GKT**$_{77–84}$, and the localization of this motif at the N-terminus of gp2 is in accordance with other phage terminases (Mitchell and Rao, 2004).

Gp3 was identified as a phage portal protein using Blast and HmmPfam searches. It shows significant homology with several phage portal proteins (Table 1). The gene encoding the portal protein is generally located immediately downstream of the terminase gene. Portal proteins are assembled typically in 12-fold symmetric ring around a central channel serving as an entry point for the translocation of the phage DNA into its head (Valpuesta and Carrascosa, 1994). Gp3 was experimentally identified as a virion protein. Gp4 is a protein of an unknown function. Most bacteriophages do not code for any protein product between the portal and prohead protease genes (Casjens, 2003; Canchaya et al., 2003). According to our phage genomes survey, the only related bacteriophage with organization similar to that of BFK20 is phi3626 (Zimmer et al., 2002). Phi3626 gp4 has no predicted function, and it does not share any sequence similarities to BFK20 gp4. An ATP binding site (P-loop) was detected in gp4 by Blast and HmmPfam searches; we speculate that gp4 is a pseudogene originating from a kinase protein family.

Gp5 is a candidate for a prohead protease, it exhibits similarity with comparable phage proteins (Table 1). An ATP/GTP-binding site motif A (P-loop) in gp5 was found using PrositeSearch. Various dsDNA phages from the Siphoviridae family share a similar capsid maturation mechanism, suggesting that the proteolytic cleavage of structural proteins during phage capsid assembly is common in many dsDNA phages and that the prohead protease is the key enzyme in the process (Liu and Mushegian, 2004). Based on the computational studies of these authors, the BFK20 prohead protease belongs to the U35 prohead protease family, according to the presence of the conserved catalytic triad **H**$_{211}$–**S**$_{262}$–**E**$_{291}$. Gp5 was experimentally identified as a virion structural protein.

Blast and HmmPfam searches predicted gp6 as the major capsid protein (MCP) with very high significance ($E = 5 \times 10^{-76}$). MCP is one the predominant proteins present in a phage virion. High level of expression for gp6 was predicted by codon usage analysis. According to the high level of sequence similarity of the prohead protease gp5 and the MCP gp6

to their homologues in HK97 and HK022, we assume that the shape and assembly process of the BFK20 capsid are analogous to those lambda-like phages, where prohead II is formed after the first 102 amino acid residues from gp5 N-terminus are cleaved off by the viral protease (Helgstrand et al., 2003; Li et al., 2005). HK97 assembly differs from most other phages in not having a scaffolding protein for assembly. It was proposed that the role of the portion of the HK97 capsid that is released upon proteolytic cleavage may compensate for the absence of the scaffolding protein (Helgstrand et al., 2003). No scaffolding protein was identified in the BFK20 genome. Based on the presence of nine strictly conserved amino acid residues, the BFK20 MCP falls within the group of HK97-like MCPs (HK97, D3, Xp10, CJW1, B1, PHI-NHI and PHI-ETA) (Helgstrand et al., 2003).

Gp8 shares significant homology with several mycobacteriophage late proteins, also without known function (Table 1). Blast predicts gp12 to be the major tail protein by its sequence similarity with several phage major tail proteins (Table 1). Sequence similarity with other late mycobacteriophage proteins without predicted function was also detected (Table 1).

Blast predicts gp15 as a minor tail protein (also known as a tape measure protein — TMP), and it was shown to be part of the BFK20 virion. Furthermore, HmmPfam identified a TMP domain at the N-terminus of gp15 and also a transglycosylase SLT (soluble lytic transglycosylase) domain. Phylogenetic analysis by Rohwer and Edwards (2002) showed a tranglycosylase domain to be one of the most frequent among phage proteins, some of which have been characterized as a tail component. Lytic transglycosylases degrade the murein component of bacterial cell wall. Many bacteriophages encode a lytic transglycosylase, which has been proposed to be involved in the entry of the phage DNA at the beginning of the infectious cycle (Lehnherr et al., 1998; Koraimann, 2003, Moak and Molineux, 2004). The presence of a transglycosylase SLT domain on BFK20 gp15 indicates that BFK20 might utilize transglycosylase activity for efficient phage DNA entry. Efficient infection of host cells is supported by transcriptional analyses of BFK20, showing that early gene transcription occurs within the first 5 min of infection (our unpublished results). N-terminal amino acid sequencing indicates that gp15 is split into a mature gp15a with 1298 AAs and gp15b with 306 AAs. The SLT domain is localized at the C-terminus of gp15a. Pedersen et al. (2000) reported a correlation between the length of the lambda TMP and the length of its tail (0.15 nm per amino acid of TMP). Using this value, we estimate the tail length of BFK20 as 195 nm, which corresponds well with previous morphological studies of BFK20 (Koptides et al., 1992).

By comparing the organization of the BFK20 genome with other tailed phage genomes, we predict that putative protein products gp16 through gp22 belong to the baseplate or tail fiber proteins. This assumption is supported only for gp16, gp18, gp21 and gp22 where several significant homologies with other late phage proteins were detected (Table 1). Gp21 is a predicted tail fiber protein, according to its similarity to several enterobacterial phages and prophages (Blast search). In the region between amino acid 105 and 277 AAs, an alanine-rich region was identified, containing 32% alanines.

Lytic proteins gp24–gp27

Most phages accomplish host lysis by at least two phage-encoded enzymes: an endolysin, which degrades the cell wall peptidoglycan and a holin, which permeabilizes the membrane. The cooperation of these proteins results in rapid destruction of the cell wall and subsequent cell lysis (Wang et al., 2000). On the BFK20 genome, we have identified two lysis genes that are clustered and transcribed in the order endolysin gp24–holin gp26.

Gp24 was identified as the phage lysin on the basis of bioinformatics and also experimentally. According to a Blast search, gp24 is similar to other autolysins/amidases. HmmPfam search revealed gp24 as a member of N-acetylmuramoyl-L-alanine amidase family. This enzyme family cleaves the amide bond between *N*-acetylmuramoyl and L-amino acids in bacterial cell walls. The lytic activity of cloned and expressed protein product gp24 was confirmed by a lysis test using *B. flavum* cells as substrate. The molecular weight of over-expressed lytic protein was estimated about 25 kDa (data not shown), which corresponds to the predicted value (Table 1).

No homology between any BFK20 protein and known holins was detected. However, holins are very diverse in primary sequence and do not share significant conserved domains. In many phage genomes, the endolysin gene is preceded or overlapped by a gene encoding a holin (Young and Bläsi, 1995). In BFK20, the gene upstream of the lysin gene is ORF23, but experimental studies and structural characteristics did not confirm gp23 to be a holin. On the basis of structural characteristics for holins, other candidates were gp26 and 27. It is known that any holin is fully functional with any endolysin (Brillard et al., 2003; Damman et al., 2000; Smith and Young, 1998). We identified the BFK20 holin gene by in vivo complementation of *λSam7* by cloned BFK20 genes. After thermal induction in a non-permissive *E. coli* strain, the λ phage lytic cycle was recovered only when BFK20 ORF26 was present (data not shown). Thus, gp26 is a functional holin.

DNA replication proteins gp29–gp46

Many bacteriophages exploit several hosts' replication proteins for genome replication (Giraldo, 2003). Bioinformatics predicted three protein products to be involved in DNA metabolism — gp41 (helicase), gp43 (replication protein) and gp44 (DNA polymerase A).

Gp41 was identified as a potential helicase. It is a member of the DEAD box helicase family (Table 1). Some phages (e.g. T4 and T5) employ separate primase and helicase proteins forming a primosome complex. Other phages use a multifunctional protein comprising both activities (e.g. gp4 of T7 and T3 and alpha protein of P4) (Deho and Ghisotti, 1999; Ilyina et al., 1992; Mosig, 1999). Studies to determine whether BFK20 also employs separate primase and helicase proteins are in progress.

Gp43 is similar to several plasmid RepA proteins and to a few bacterial helicase-like proteins (Table 1). RepA-like

proteins typically contain an ATP-binding site and are responsible for plasmid replication. Gp43 contains an ATP/GTP-binding motif A (P-loop). Gp43 is apparently another protein participating in a phage DNA replication, operating, for example, as primase or second helicase.

We predicted gp44 as a DNA polymerase. This finding is highly supported by Blast ($E = 5.0 \times 10^{-18}$), which revealed homology to several bacteriophage and bacterial A family DNA polymerases (Table 1). Surprisingly, gp44 is more closely related to e.g. enterobacterial DNA polymerases than to *Corynebacterium* spp. DNA polymerases.

Other early proteins include several that probably play a role in DNA metabolism and that show significant homologies with other early phage proteins. Gp29 is a protein of unknown function; a PSI-Blast search revealed similarity to other bacterial and phage proteins, some belonging to the endonuclease VII family (Table 1). The position of gp29 in the early gene region, together with its similarities to the endonuclease VII family, points to a potential role in BFK20 DNA replication. Gp32 contains a MerR family regulatory protein signature at its N-terminal end. The MerR-type HTH domain is a DNA-binding winged helix–turn–helix domain of about 70 residues present in the MerR family of transcriptional regulators. Thus, BFK20 gp32 is a putative transcriptional regulator. Gp37 is predicted as another putative transcriptional regulatory protein with a helix–turn–helix motif; gp37 also contains an adenylate cyclase associated domain (Table 1). Gp39 is a protein with unknown function, the first 153 AAs contain a glutamine-rich (43%) and proline-rich (20%) region. Several repeats were identified at the amino acid level near the C-terminus of gp39; repeats are found in 12–14 AA shifts. At the DNA level, in the region 28,850–29,250 bp (corresponding to the C-terminal part of gp39), several direct and inverted repeats were also identified. The largest direct perfect repeat is 92 bp. These repeated sequences resulted in many false matches using Blast and HmmPfam searches. Gp40 and gp42 are both early proteins with unknown functions with homology to several putative phage proteins — also with unknown function (Table 1).

Regulatory proteins gp47–gp55

The life cycle of bacteriophages is altered by phage-encoded regulatory proteins. Those include transcription repressors and antirepressors, which enable switching on and off expression of particular groups of genes. These proteins, together with host-encoded proteins, regulate the development phases of a phage. We identified six transcriptional regulators: four (gp32, gp37, gp47 and gp48) lie within early genes and two (gp52 and gp54) within late genes. Gp47, gp48, gp52 and gp54 are all in the region where transcription of both early and late genes is initiated.

Gp47 contains Pfam transcriptional regulatory motif. This Pfam family covers small proteins of 12 to 18 kDa that seem to contain a signal sequence. Gp48 is a member of Pfam bacterial regulatory GntR family of transcription factors, which contain an N-terminal helix–turn–helix region. Gp52 is a member of lactococcal phage putative transcription regulator family. This family contains a number of putative transcription repressor proteins found in several lactococcal bacteriophages. Gp54 contains a lambda-repressor HTH signature and is another potential regulatory protein.

In this region of the BFK20 genome, we also detected two non-regulatory proteins. Gp53 is a glutaredoxin-like protein (Table 1); glutaredoxins, also known as thioltransferases, and thioredoxins have important roles in the DNA metabolism of diverse viruses (Gvakharia et al., 1996). Glutaredoxin-like proteins have been identified in a related group of mycobacteriophages (Pedulla et al., 2003). Gp55 is predicted HNH endonuclease with an $\mathbf{H}_{52}$–$\mathbf{N}_{68}$–$\mathbf{N}_{77}$ motif. Substitution of histidine with asparagine in the third position is present in some members of the HNH family (Dalgaard et al., 1997).

Thirty two proteins are without predicted function (Table 1), showing no or only low similarity with other known proteins or protein families. Eighteen are less than 100 amino acids long, thus it was difficult to identify specific motifs. Further experimental investigation will be necessary to verify if the corresponding ORFs are functional genes or pseudogenes.

Bacteriophage BFK20 genome similarities

Genome mosaicism is a property first described for lambdoid phages following electron microscopy of DNA–DNA heteroduplexes. However, modular theory of phage evolution can be applied to all bacteriophages. This theory hypothesizes that modules of sequence, consisting of genes or group of genes, exchange by non-homologous recombination at random locations. Only non-lethal recombinants will persist in subsequent generations, possibly providing a phage with a new or enhanced functionality (Hendrix, 2002, Hendrix et al., 2003).

Our bioinformatics analyses indicated quite extensive homology between BFK20 and other bacteriophages (particularly those using *Mycobacterium* spp. as hosts) and to regions of *Corynebacterium* spp. genomes. Most notable are the similarities between the BFK20 genome with the mycobacteriophage Rosebush and prophages from *C. efficiens* YS-314 and *C. diphtheriae* NCTC 13129 and *C. glutamicum* strain specific island 8, strain R. We used these phage/prophage genome sequences, together with BFK20 genome, for multi-dot plot analysis (Fig. 3). While early genes, particularly ORF40–ORF44 are very similar to Rosebush DNA replication genes, late genes (structural and lytic) are more related to the *Corynebacterium* spp. prophages. Within the late gene group, we find great similarity between ORF1–ORF15 (terminase, capsid, major tail protein) to the *C. diphtheriae* prophage. *C. glutamicum* and *C. efficiens* prophages share greater similarities with ORF15–ORF24 (minor tail, lysin), and, for the *C. efficiens* prophage, the region of similarity extends through ORF27. At the protein level, all phage capsid proteins – portal protein (gp3), prohead protease (gp5) and major capsid protein (gp6) – are homologous to the capsid proteins of lambda-like phages (HK022 and HK97). Evidence for common ancestors of capsid proteins confirmed that the groups of genes whose protein product closely interact are usually transferred horizontally as a one module.

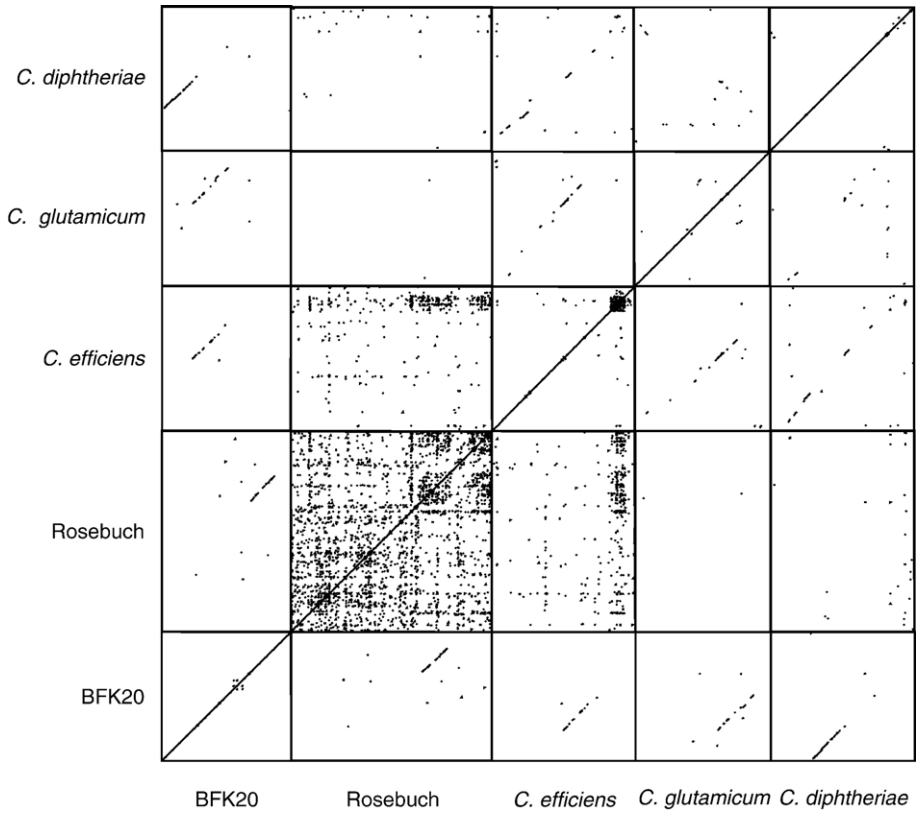


Fig. 3. Multiple DotPlot of BFK20 genome sequence and related phage/prophage sequences. For analysis, the whole genome sequences of corynephage BFK20 (42,968 bp) and mycobacteriophage Rosebush (EMBL Acc. AY129334, 67,480 bp) were used together with probable prophage sequences extracted from bacterial genomes of *Corynebacterium diphtheriae* NCTC 13129 (EMBL Acc. BX248353; genome position 152,718–200,798 bp), *C. glutamicum* strain specific island 8, strain: R (EMBL Acc. AB193035) and *Corynebacterium efficiens* YS-314 (EMBL Acc. BA000035; genome position 874,694–922,705 bp). Windows size of 150 and threshold value of 50 were used as parameters for DotmatcherN program (EMBOSS package).

Analyses at both the DNA and at the protein level support the theory that phage genomes are mosaics with respect to each other. While the BFK20 capsid protein is mostly closely related to the lambda-like group of enterophages, regions containing other structural and replication genes are more homologous to phages and prophages from the *Corynebacterineae* taxonomical suborder.

Materials and methods

Phage propagation and isolation of BFK20 genomic DNA

Corynephage BFK20 was propagated on *B. flavum* CCM 251 (hse^{-}, Aec^{r}) L-lysine producer. High titer lysates preparation and phage DNA isolation were performed as previously described (Koptides et al., 1992). The phage particles were purified on CsCl step gradient according to Sambrook et al. (1989). The obtained phage DNA was subcloned into pBSSKII(+) and pBSKSII(+) plasmids (Stratagene, La Jolla, CA).

DNA manipulations and sequencing strategy

The BFK20 genome library was prepared by cloning of *Bgl*II and/or *Bgl*II/*Mlu*I generated fragments and by shotgun cloning of *Bam*HI fragments into pBluescript plasmids. The recombinant plasmids, with fragments longer than 1 kb, were digested with *Sac*I and *Xba*I and treated with exonuclease III (Promega, Madison, USA) to generate a series of unidirectional deleted subclones. Isolated plasmids were sequenced using T3 and T7 primers. The sequencing was performed by using the Thermo Sequenase fluorescent-labeled primer cycle sequencing Kit (Amersham Biosciences, Essex, UK) on a model A.L.F. DNA sequencer (Pharmacia-LKB), on Vistra DNA Sequencer (Amersham Biosciences, Essex, UK) and using an 8-column capillary Beckman CEQ sequencer with reagents and methods recommended by Beckman (Beckman Instruments).

Determination of BFK20 cos sites

The sequence of *cos* site of BFK20 was determined by a sequence run-off experiment using the oligonucleotide COS1 (5′-CAACTCGAGCCACCATGAAG-3′) and COS2 (5′-TTGGTAAAGTGAGGCCCAAG-3′), which are situated 277 bp distal to either end of the linear BFK20 genome. The purified phage BFK20 DNA was used as a sequencing template. The nucleotide sequences were determined using an 8-column capillary ABI 3100-Avant Genetic Analyser sequencer with reagents and methods recommended by Applied Biosystems.

Analysis and amino terminal sequencing of BFK20 structural proteins

Two-dimensional gel electrophoresis was performed according to Garrels (1979). Isoelectric focusing gels contained 2% of ampholines, pH range 3–10 (Bio-Rad). The second dimension was performed on 13% polyacrylamide gels according to Laemmli (1970). Gels were silver-stained for detection of the proteins and dried. The proteins were transferred to a PVDF membrane and subjected to an automated Edman degradation sequencing.

Western blots

Electrotransfer of gels to the Hybond P, PVDF transfer membrane (Amersham Pharmacia Biotech) was performed using a blotting apparature (220 mA, 2 h). PVDF blots were rinsed in deionized water then soaked in pure methanol for a few seconds and stained with Coomassie Blue R-250 (0.1% in 1% acetic acid). After destaining procedure, the PVDF blots were rinsed in deionized water and dried at a room temperature. Protein spots were cut out and marked according to the p*I* and Mw markers.

N-terminal amino acid sequences were determined using the "Procise-Protein Sequencing System" [PE Applied Biosystems, 491 Protein Sequencer] by Edman degradation (the program PL PVDF Protein was used). N-terminal amino acid was specifically reacted with phenylisothiocyanate (PITC). This derivatized amino acid was then selectively removed, leaving the rest of peptide chain intact. The resulting PTH amino acids were analyzed sequentially [by RP HPLC] to determine the amino acid sequence of the protein. Blast search of the obtained sequences against the BFK20 genome sequence was performed (tblastn algorithm).

ORFs prediction and genome annotation

ORFs were predicted using Staden Package program Spin (Staden, 1996), based on base preferences, codon preferences, author test, base bias and the presence of START and STOP codons. For those methods requiring Codon Usage Table, the one for *C. glutamicum* was used.

Ribosome binding sites (RBS) were predicted using TIGR RBSfinder (http://www.tigr.org/software/). As an RBS consensus, we used the AGGAG sequence that is complement to the sequence motif located on the 3′ end of 16S rRNA of *Corynebacterium* spp. small ribosomal subunit.

Transcriptional terminator was predicted using the Terminator tool from GCG package, the Wisconsin Sequence Analysis Package version 10, using algorithm of Brendel and Trifonov (1984).

Codon usage analysis was performed using INCA 2.0 software (Supek and Vlahovicek, 2004). Complete set of *C. glutamicum* ATCC 13032 was obtained from NCBI ftp server (ftp://ftp.ncbi.nih.gov/genomes/Bacteria).

Primary databases search was done by NCBI Blast server (Altschul et al., 1997). Both blastn and blastp programs and corresponding non-redundant databases were used, depending on the nature of analysis. Only those results with the E-value smaller than 10 were considered. If standard Blast search was not sensitive enough, PSI-Blast with up to 5 itineration was used.

Secondary databases search was performed against Prosite (Hulo et al., 2004), Prints (Attwood et al., 2003) and Pfam (Bateman et al., 2004) databases. Corresponding tools PrositeSearch (EMBOSS package software, Rice et al., 2000), Pscan (EMBOSS package software, Rice et al., 2000), FPScan (Scordis et al., 1999) and HmmPfam (Durbin et al., 1998) were used for analysis.

Membrane spanning regions in proteins were predicted using Tmap tool (EMBOSS package software, Rice et al., 2000).

Multiple sequence alignments were generated by employing the ClustalX program (Thompson et al., 1997), using standard parameters.

DotPlot were constructed with Dotmatcher tool from EMBOSS package (Rice et al., 2000). Windows size of 150 and threshold value of 50 were used as parameters.

Nucleotide sequence accession number

The complete nucleotide sequence of phage BFK20 was deposited in the EMBL database under accession number AJ278322.

Acknowledgments

This work was supported by VEGA grant 2/5068/25 from Slovak Academy of Sciences, APVT-51-025004 grant from Science and Technology Assistance Agency and in part by project no. AV0Z50520514 awarded by the Academy of Sciences of the Czech Republic. We would like to thank Dr. Miroslav Kubeš and Mgr. Robert Ihnatko (Institute of Virology SAS, Bratislava, Slovakia) for performing two-dimensional gel electrophoresis and Western blots and Dr. Zdenek Voburka (Institute of Organic chemistry and Biochemistry, CAS, Prague, Czech Republic) for N-terminal sequencing. The authors are grateful to Dr. Vladimir Kery (Systems Biology PNNL, WA, U. S.A.) for critical reading of the manuscript.

Appendix A. Supplementary data

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.virol.2005.12.010.

References

Ackermann, H.W., 2003. Bacteriophage observations and evolution. Res. Microbiol. 154, 245–251.

Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W., Lipman, D.J., 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 25, 3389–3402.

Attwood, T.K., Bradley, P., Flower, D.R., Gaulton, A., Maudling, N., Mitchell, A.L., Moulton, G., Nordle, A., Paine, K., Taylor, P., 2003. PRINTS and its automatic supplement, prePRINTS. Nucleic Acids Res. 31, 400–402.

Bateman, A., Coin, L., Durbin, R., Finn, R.D., Hollich, V., Griffiths-Jones, S.,

Khanna, A., Marshall, M., Moxon, S., Sonnhammer, E.L.L., Studholme, D.J., Yeats, C., Eddy, S.R., 2004. The Pfam protein families database. Nucleic Acids Res. 32, D138–D141 (Database issue).

Batt, C.A., Follettie, M.T., Shen, H.K., Yeh, P., Sinskey, A.J., 1985. Genetic engineering of coryneform bacteria. Trends Biotechnol. 3, 305–310.

Brendel, V., Trifonov, E.N., 1984. A computer algorithm for testing potential prokaryotic terminators. Nucleic Acids Res. 12, 4411–4427.

Brillard, J., Boyer-Giglio, M.H., Boemare, N., Givaudan, A., 2003. Holin locus characterisation from lysogenic *Xenorhabdus nematophila* and its involvement in *Escherichia coli* SheA haemolytic phenotype. FEMS Microbiol. Lett. 218, 107–113.

Brüssow, H., Desiere, F., 2001. Comparative phage genomics and the evolution of Siphoviridae: insights from dairy phages. Mol. Microbiol. 39, 213–222.

Buck, G.A., Groman, N.B., 1981. Physical mapping of beta-converting and gamma-nonconverting corynebacteriophage genomes. J. Bacteriol. 148, 131–142.

Canchaya, C., Proux, C., Fournous, G., Bruttin, A., Brussow, H., 2003. Prophage genomics. Microbiol. Mol. Biol. Rev. 67, 238–276.

Catalano, C.E., Cue, D., Feiss, M., 1995. Virus DNA packaging: the strategy used by phage lambda. Mol. Microbiol. 16, 1075–1086.

Casjens, S., 2003. Prophages and bacterial genomics: what have we learned so far? Mol. Microbiol. 49, 277–300 (Review).

Chandry, P.S., Moore, S.C., Davidson, B.E., Hillier, A.J., 1994. Analysis of the *cos* region of the *Lactococcus lactis* bacteriophage sk1. Gene 138, 123–126.

Dalgaard, J.Z., Klar, A.J., Moser, M.J., Holley, W.R., Chatterjee, A., Mian, I.S., 1997. Statistical modeling and analysis of the LAGLIDADG family of site-specific endonucleases and identification of an intein that encodes a site-specific endonuclease of the HNH family. Nucleic Acids Res. 25, 4626–4638.

Damman, C.J., Eggers, C.H., Samuels, D.S., Oliver, D.B., 2000. Characterisation of *Borrelia burgdorferi* Bly A and Bly B proteins: a prophage encoded holin-like system. J. Bacteriol. 182, 6791–6797.

Deho, G., Ghisotti, D., 1999. P4 phage (satellites). In: Granoff, A., Webster, R. (Eds.), Encyclopedia of Virology. Academic Press, pp. 1094–1104.

Duffy, C., Feiss, M., 2002. The large subunit of bacteriophage lambda's terminase plays a role in DNA translocation and packaging termination. J. Mol. Biol. 316, 547–561.

Durbin, R., Eddy, S., Krogh, A., Mitchison, G., 1998. Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids. Cambridge Univ. Press.

Ganyu, A., Csiszovszki, Z., Ponyi, T., Kern, A., Buzás, Z., Orosz, L., Papp, P.P., 2005. Identification of cohesive ends and genes encoding the terminase of phage 16-3. J. Bacteriology 187, 2526–2531.

Garrels, J.I., 1979. Two-dimensional gel electrophoresis and computer analysis of proteins synthesized by clonal cell lines. J. Biol. Chem. 254, 7961–7977.

Gilakjan, Z.A., Kropinski, A.M., 1999. Cloning and analysis of the capsid morphogenesis genes of *Pseudomonas aeruginosa* bacteriophage D3: another example of protein chain mail? J. Bacteriol. 181, 7221–7227.

Giraldo, R., 2003. Common domains in the initiators of DNA replication in Bacteria, Archaea and Eukarya: combined structural, functional and phylogenetic perspectives. FEMS Microbiol. Rev. 26, 533–554.

Gvakharia, B.O., Hanson, E., Koonin, E.K., Mathews, C.K., 1996. Identification of a second functional glutaredoxin encoded by the bacteriophage T4 genome. J. Biol. Chem. 271, 15307–15310.

Halgasova, N., Majtan, T., Ugorcakova, J., Timko, J., Bukovska, G., 2005. Resistance of corynebacterial strains to infection and lysis by corynephage BFK20. J. Appl. Microbiol. 98, 184–192.

Helgstrand, C., Wikoff, W.R., Duda, R.L., Hendrix, R.W., Johnson, J.E., Liljas, L., 2003. The refined structure of a protein catenane: the HK97 bacteriophage capsid at 3.44 A resolution. J. Mol. Biol. 334, 885–899.

Hendrix, R.W., 2002. Bacteriophages: evolution of the majority. Theor. Popul. Biol. 61, 471–480.

Hendrix, R.W., Hatfull, G.F., Smith, M.C., 2003. Bacteriophages with tails: chasing their origins and evolution. Res. Microbiol. 154, 253–257.

Hermann, T., 2003. Industrial production of amino acids by coryneform bacteria. J. Biotechnol. 104, 155–172.

Hulo, N., Sigrist, C.J., Le Saux, V., Langendijk-Genevaux, P.S., Bordoli, L., Gattiker, A., De Castro, E., Bucher, P., Bairoch, A., 2004. Recent improvements to the PROSITE database. Nucleic Acids Res. 32, D134–D137 (Database issue).

Ikeda, M., Nakagawa, S., 2003. The *Corynebacterium glutamicum* genome: features and impacts on biotechnological processes. Appl. Microbiol. Biotechnol. 62, 99–109.

Ilyina, T.V., Gorbalenya, A.E., Koonin, E.V., 1992. Organization and evolution of bacterial and bacteriophage primase-helicase systems. J. Mol. Evol. 34, 351–357.

Kalinowski, J., Bathe, B., Bartels, D., Bischoff, N., Bott, M., Burkovski, A., Dusch, N., Eggeling, L., Eikmanns, B.J., Gaigalat, L., Goesmann, A., Hartmann, M., Huthmacher, K., Kramer, R., Linke, B., McHardy, A.C., Meyer, F., Mockel, B., Pfefferle, W., Puhler, A., Rey, D.A., Ruckert, C., Rupp, O., Sahm, H., Wendisch, V.F., Wiegrabe, I., Tauch, A., 2003. The complete *Corynebacterium glutamicum* ATCC 13032 genome sequence and its impact on the production of L-aspartate-derived amino acids and vitamins. J. Biotechnol. 104, 5–25.

Kaneko, J., Kimura, T., Narita, S., Tomita, T., Kamio, Y., 1998. Complete nucleotide sequence and molecular characterization of the temperate staphylococcal bacteriophage fPVL carrying Panton–Valentine leukocidin genes. Gene 215, 57–67.

Karlin, S., Campbell, A.M., Mrazek, J., 1998a. Comparative DNA analysis across diverse genomes. Annu. Rev. Genet. 32, 185–225.

Karlin, S., Mrazek, J., Campbell, A.M., 1998b. Codon usages in different gene classes of the *Escherichia coli* genome. Mol. Microbiol. 29, 1341–1355.

Kato, F., Yoshimi, M., Araki, K., Motomura, Y., Matsufune, Y., Nobunaga, H., Murata, A., 1984. Screening of bacteriocins in amino acid or nucleic acid producing bacteria and related species. Agric. Biol. Chem. 48, 193–200.

Koptides, M., Barak, I., Sisova, M., Baloghova, E., Ugorcakova, J., Timko, J., 1992. Characterization of bacteriophage BFK20 from *Brevibacterium flavum*. J. Gen. Microbiol. 138, 1387–1391.

Koptides, M., Ugorcakova, J., Baloghova, E., Bukovska, G., Timko, J., 1994. Characterization and sequence analysis of the F2 promoter from corynephage BFK20. Acta Virol. 38, 223–228.

Koraimann, G., 2003. Lytic transglycosylases in macromolecular transport systems of Gram-negative bacteria. Cell. Mol. Life Sci. 60, 2371–2388.

Krämer, R., 1996. Genetic and physiological approaches for the production of amino acids. J. Biotechnol. 45, 1–21.

Laemmli, U.K., 1970. Cleavage of structural proteins during the assembly of the head of bacteriophage T4. Nature 227, 680–685.

Labrie, S., Moineau, S., 2002. Complete genomic sequence of bacteriophage ul36: demonstration of phage heterogeneity within the P335 quasi-species of lactococcal phages. Virology 296, 308–320.

Lehnherr, H., Hansen, A.M., Ilyina, T., 1998. Penetration of the bacterial cell wall: a family of lytic transglycosylases in bacteriophages and conjugative plasmids. Mol. Microbiol. 30, 454–457.

Le Marrec, C., Moreau, S., Loury, S., Blanco, C., Trautwetter, A., 1996. Genetic characterization of site-specific integration functions of phi AAU2 infecting "*Arthrobacter aureus*" C70. J. Bacteriol. 178, 1996–2004.

Li, Y., Conway, J.F., Cheng, N., Steven, A.C., Hendrix, R.W., Duda, R.L., 2005. Control of virus assembly: HK97 "Whiffleball" mutant capsids without pentons. J. Mol. Biol. 348, 167–182.

Liebl, W., Ehrmann, M., Ludwig, W., Schleifer, K.H., 1991. Transfer of Brevibacterium divaricatum DSM 20297T, '*Brevibacterium flavum*' DSM 20411, '*Brevibacterium lactofermentum*' DSM 20412 and DSM 1412, and *Corynebacterium lilium* DSM 20137T to *Corynebacterium glutamicum* and their distinction by rRNA gene restriction patterns. Int. J. Syst. Bacteriol. 41, 255–260.

Lillehaug, D., Lindqvist, B.H., Birkland, N.K., 1993. Characterization of fLC3, a *Lactococcus lactis* subsp. *cremoris* temperate bacteriophage with cohesive single-stranded DNA ends. Appl. Environ. Microbiol. 57, 3206–3211.

Liu, J., Mushegian, A., 2004. Displacements of prohead protease genes in the late operons of double-stranded-DNA bacteriophages. J. Bacteriol. 186, 4369–4375.

Lowther, W.T., Matthews, B.W., 2000. Structure and function of the methionine aminopeptidases. Biochim. Biophys. Acta 1477, 157–167.

Mahanivong, C., Boyce, J.D., Davidson, B.E., Hillier, A.J., 2001. Sequence analysis and molecular characterization of the *Lactococcus lactis* temperate bacteriophage BK5-T. Appl. Environ. Microbiol. 67, 3564–3576.

Maluf, N.K., Yang, Q., Catalano, C.E., 2005. Self-association properties of the bacteriophage λ terminase holoenzyme: implications for the DNA packaging motor. J. Mol. Biol. 347, 523–542.

Martin, J.F., Barreiro, C., Gonzalez-Lavado, E., Barriuso, M., 2003. Ribosomal RNA and ribosomal proteins in corynebacteria. J. Biotechnol. 104, 41–53.

Mitchell, M.S., Rao, V.B., 2004. Novel and deviant Walker A ATP-binding motifs in bacteriophage large terminase-DNA packaging proteins. Virology 321, 217–221.

Moak, M., Molineux, I.J., 2004. Peptidoglycan hydrolytic activities associated with bacteriophage virions. Mol. Microbiol. 51, 1169–1183.

Moreau, S., Leret, V., Le Marrec, C., Varangot, H., Ayache, M., Bonnassie, S., Blanco, C., Trautwetter, A., 1995. Prophage distribution in coryneform bacteria. Res. Microbiol. 146, 493–505.

Moreau, S., Blanco, C., Trautwetter, A., 1999. Site-specific integration of corynephage phi16: construction of an integration vector. Microbiology 145, 539–548.

Mosig, G., 1999. T4-like phages (Myoviridae). In: Granoff, A., Webster, R. (Eds.), Encyclopedia of Virology. Academic Press, pp. 1706–1716.

Nampoothiri, M., Pandey, A., 1998. Genetic tuning of coryneform bacteria for the overproduction of amino acids. Process Biochem. 33, 147–161.

Park, Y.H., Hori, H., Suzuki, K.I., Osawa, S., Komagata, K., 1987. Phylogenetic analysis of the coryneform bacteria by 5S rRNA sequences. J. Bacteriol. 169, 1801–1806.

Patek, M., Ludvik, J., Benada, O., Hochmannova, J., Nešvera, J., Krumphanzl, V., Buèko, M., 1985. New bacteriophage-like particles in *Corynebacterium glutamicum*. J. Virol. 140, 360–363.

Pedersen, M., Ostergaard, S., Bresciani, J., Vogensen, F.K., 2000. Mutational analysis of two structural genes of the temperate lactococcal bacteriophage TP901-1 involved in tail length determination and baseplate assembly. Virology 276, 315–328.

Pedulla, M.L., Ford, M.E., Houtz, J.M., Karthikeyan, T., Wadsworth, C., Lewis, J.A., Jacobs-Sera, D., Falbo, J., Gross, J., Pannunzio, N.R., Brucker, W., Kumar, V., Kandasamy, J., Keenan, L., Bardarov, S., Kriakov, J., Lawrence, J.G., Jacobs Jr., W.R., Hendrix, R.W., Hatfull, G.F., 2003. Origins of highly mosaic mycobacteriophage genomes. Cell 113, 171–182.

Rice, P., Longden, I., Bleasby, A., 2000. EMBOSS: the European molecular biology open software suite. Trends Genet. 16, 276–277.

Roberts, M.D., Martin, N.L., Kropinski, A.M., 2004. The genome and proteome of coliphage T1. Virology 318, 245–266.

Rohwer, F., Edwards, R., 2002. The Phage Proteomic Tree: a genome-based taxonomy for phage. J. Bacteriol. 184, 4529–4535.

Romero, P., Lopez, R., Garcia, E., 2004. Genomic organization and molecular analysis of the inducible prophage EJ-1, a mosaic myovirus from an atypical pneumococcus. Virology 322, 239–252.

Rubinchik, S., Parris, W., Gold, M., 1994. The in vitro endonuclease activity of gene product A, the large subunit of the bacteriophage lambda terminase, and its relationship to the endonuclease activity of the holoenzyme. J. Biol. Chem. 269, 13575–13585.

Sambrook, J., Fritsch, E.F., Maniatis, T., 1989. Molecular Cloning: A Laboratory Manual, 2nd ed. Cold Spring Harbor Laboratory Press, New York, NY.

van Sinderen, D., Karsens, H., Kok, J., Terpstra, P., Ruiters, M.H.J., Venema, G., Nauta, A., 1996. Sequence analysis and molecular characterization of the temperate lactococcal bacteriophage rlt. Mol. Microbiol. 19, 1343–1355.

Scordis, P., Flower, D.R., Attwood, T.K., 1999. FingerPRINTScan: intelligent searching of the PRINTS motif database. Bioinformatics 15, 799–806.

Sharp, P.M., Li, W.H., 1987. The codon adaptation index—A measure of directional synonymous codon usage bias, and its potential applications. Nucleic Acids Res. 15, 1281–1295.

Smith, D.L., Young, R., 1998. Oligohistidine tag mutagenesis of the λ holin gene. J. Bacteriol. 180, 4199–4211.

Smith, M.C., Burns, R.N., Wilson, S.E., Gregory, M.A., 1999. The complete genome sequence of the *Streptomyces* temperate phage straight phiC31: evolutionary relationships to other viruses. Nucleic Acids Res. 27, 2145–2155.

Sonnen, H., Schneider, J., Kutzner, H.J., 1990a. Characterization of phi GA1, an inducible phage particle from *Brevibacterium flavum*. J. Gen. Microbiol. 136, 567–571.

Sonnen, H., Schneider, J., Kutzner, H.J., 1990b. Corynephage Cog, a virulent bacteriophage of *Corynebacterium glutamicum*, and its relationship to phi GA1, an inducible phage particle from *Brevibacterium flavum*. J. Gen. Virol. 71, 1629–1633.

Staden, R., 1996. The staden sequence analysis package. Mol. Biotechnol. 5, 233–241.

Supek, F., Vlahovicek, K., 2004. INCA: synonymous codon usage analysis and clustering by means of self-organizing map. Bioinformatics 20, 2329–2330.

Thompson, J.D., Gibson, T.J., Plewniak, F., Jeanmougin, F., Higgins, D.G., 1997. The ClustalX windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. Nucleic Acids Res. 24, 4876–4882.

Trautwetter, A., Blanco, C., 1988. Isolation and preliminary characterization of twenty bacteriophages infecting either *Brevibacterium* or *Arthrobacter* strains. Appl. Environ. Microbiol. 54, 1466–1471.

Trautwetter, A., Blanco, C., Sicard, A.M., 1987a. Structural characteristics of the *Corynebacterium lilium* bacteriophage CL31. J. Virol. 61, 1540–1545.

Trautwetter, A., Blanco, C., Bonnassie, S., 1987b. Characterization of the corynebacteriophage CG33. J. Gen. Microbiol. 133, 2945–2952.

Valpuesta, J.M., Carrascosa, J.L., 1994. Structure of viral connectors and their function in bacteriophage assembly and DNA packaging. Q. Rev. Biophys. 27, 107–155.

Wang, I.N., Smith, D.L., Young, R., 2000. HOLINS: the protein clock of bacteriophage infection. Annu. Rev. Microbiol. 54, 799–825.

Woese, C.R., 1987. Bacterial evolution. Microbiol. Rev. 51, 221–271.

Xu, J., Hendrix, R.W., Duda, R.L., 2004. Conserved translational frameshift in dsDNA bacteriophage tail assembly genes. Mol. Cell 16, 11–21.

Young, R., Bläsi, U., 1995. Holins: form and function in bacteriophage lysis. FEMS Microbiol. Rev. 17, 191–205.

Yuzenkova, J., Nechaev, S., Berlin, J., Rogulja, D., Kuznedelov, K., Inman, R., Mushegian, A., Severinov, K., 2003. Genome of *Xanthomonas oryzae* bacteriophage Xp10: an odd T-odd phage. J. Mol. Biol. 330, 735–748.

Zimmer, M., Scherer, S., Loessner, M.J., 2002. Genomic analysis of *Clostridium perfringens* bacteriophage phi3626, which integrates into guaA and possibly affects sporulation. J. Bacteriol. 184, 4359–4368.